Working Paper Series
ISSN 1170-487X

INTERACTIVE DOCUMENT SUMMARISATION

By Steve Jones, Stephen Lundy and Gordon W. Paynter

Working Paper: 01/1
February 2001

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Interactive Document Summarisation

Steve Jones Stephen Lundy Gordon W. Paynter

Department of Computer Science
University of Waikato
Private Bag 3105, Hamilton, New Zealand
00 64 7 838 4021
{stevej, paynter}@cs.waikato.ac.nz

ABSTRACT

This paper describes the Interactive Document Summariser (IDS), a dynamic document summarisation system, which can help users of digital libraries to access on-line documents more effectively. IDS provides dynamic control over summary characteristics, such as length and topic focus, so that changes made by the user are instantly reflected in an on-screen summary. A range of 'summary-in-context' views support seamless transitions between summaries and their source documents. IDS creates summaries by extracting keyphrases from a document with the Kea system, scoring sentences according to the keyphrases that they contain, and then extracting the highest scoring sentences. We report an evaluation of IDS summaries, in which human assessors identified suitable summary sentences in source documents, against which IDS summaries were judged. We found that IDS summaries were better than baseline summaries, and identify the characteristics of Kea keyphrases that lead to the best summaries.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *user issues*. I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis*. H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering*.

General Terms

Algorithms, Performance, Experimentation.

Keywords

dynamic document summarisation, sentence extraction, automatic keyphrasing, evaluation

1. INTRODUCTION

As information retrieval systems like digital libraries and Web search engines become more ubiquitous, the efficacy with which they identify information that matches a user's needs increases. Query interfaces, storage techniques and retrieval algorithms are developing rapidly. However, no matter how effective each of these attributes of a digital library, it is likely that not all items deemed useful by the system will be of interest to the user.

Therefore, the onus falls upon the user to discriminate between interesting and non-interesting items, such as documents. Whereas this document provides a number of useful cues regarding its content—a title, the authors, an abstract, subject descriptors, a list of keywords—many on-line documents do not. Often, the user receives little support from the system to identify useful documents. For example, web search engines present long lists of query results containing little more than the title of a document or Web page. In fact, users need to be supported in making judgements that are far subtler than classifying documents as useful/non-useful. For example, segments of a document rather than the whole might meet their needs, and need to be located and considered. The difference between two ostensibly similar documents may need to be determined to select the most appropriate. Generally, users must incur the cost of reading a document to make these complex judgements. Furthermore, a document may be too long to quickly assess, either wholly or in part.

A promising way to address this problem is to produce summaries of documents that may be provided in result lists, as intermediaries between result lists and entire documents, or as surrogates for entire documents. Of course, manual production of summaries is a time-consuming and expensive task, particularly on the scale required by most information providers. Automated summarisation tools can overcome these limitations.

Unfortunately, automated summarisation tools conventionally force providers and users to remain passive, with little control over the summaries that are produced. Users and providers cannot tailor summaries to their requirements, in respect of their length and content. Further, the summaries retain little context from, or means of access into, the full document from which they were produced.

In this paper we present the Interactive Document Summariser (IDS), a system that supports dynamic control over the production of document summaries. IDS allows users to tailor the length and content of a summary, seeing changes in real-time, as they amend summary attributes. It also provides a number of visualisations of the summary, to support interpretation in the context of the entire document. IDS produces summaries by identifying and extracting sentences that best reflect what a document is about, and does this based on sets of automatically extracted keyphrases.

We believe that IDS can be useful in a number of ways. First it can help information providers to produce tailored summaries for presentation to users. Second, it can support document authors in creating summaries of their texts. Third, when embedded into information access interfaces, it can help users to more effectively determine whether long on-line documents are useful to them.

This paper is organised as follows: in the next section we present a brief overview of document summarisation techniques. Following this, we describe the few existing approaches to interactive summarisation. We then describe the IDS systems, detailing both its interface and underlying summarisation mechanisms, and go on to report an initial evaluation of IDS summaries. Finally we present our conclusions and outline avenues for future work.

2. DOCUMENT SUMMARISATION

Summarisation systems generate concise descriptions of the content of a document, mainly either by *abstraction* or *extraction*. The goal of abstraction is to produce summaries that read as coherently as text produced by humans. This is difficult to achieve with current natural language processing techniques [10]. Consequently extraction techniques have formed the primary focus of summarisation research. The goal is to identify a set of text segments that reflect the content of a document. A number of granularities of segment have been suggested, ranging from keywords and phrases, [8, 14, 21], to paragraphs [18]

Sentences are commonly chosen as the target segments to extract [8, 10, 15, 16, 20]. The sentence extraction process is essentially as follows: apply a mechanism to allocate a score to each sentence in the text, rank all sentences by decreasing score, and finally select the N highest scoring sentences to form the summary. N may be an absolute value, or expressed as some fraction of the original document. This approach is rather simpler than abstraction, but suffers from unresolved co-references, anaphora and so on.

Nevertheless, many scoring heuristics have been suggested, often weighting multiple attributes of a sentence to produce a score. Some attributes are simple to compute, such as sentence length (to favour longer sentences), and whether a sentence includes certain cue phrases like “In conclusion” or “In this paper”. Location in the text is often used to favour sentences that are closer to the start of a document. Structural information, such as section headings may be identified, so that initial sentences in sections can be weighted more strongly.

Statistical analysis techniques can be used to identify important words or phrases in a document, and sentences can then be scored based on the occurence of such words and phrases within them. Similarly, lexical connectivity (commonality of terms) between sentences can be calculated and used for scoring purposes. Sentence attributes can be used individually or in combination. Lin [16], for example, presents a scoring heuristic using ten attributes in combination. Attributes are often weighted in a heuristic manner, but some research has treated sentence extraction as a learning problem [15, 16, 20]. In this approach, training material exemplifies the nature of desired summaries by providing document-summary pairs. From this a classification model can be built, and applied to previously unseen documents.

Assessing the quality of produced summaries is a difficult task. A range of measures has been used, with wide varieties of test corpora, and it is consequently difficult to characterise the state of the art. Standard information retrieval measures of precision and recall can indicate the proportion of ‘good’ sentences in a summary, and how well a summary covers all ‘good’ sentences in a document. Normalised recall [10] takes into account the fact that a summary is unlikely to be able to return all ‘good’ sentences from a document. Combined precision and recall can be represented using the F-score (or normalised F-score), showing summariser performance at different levels of compression of the source document.

A summariser may be used to produce summaries off-line, or in response to queries issued by users. The latter case emphasises that there may be no ideal single summary for a document. Indeed, in the case of query-weighted summaries some inputs to a summariser (the query terms) must be dynamically specified. A range of other inputs may be specified to bias the output, such as summary length, minimum sentence length and so on. These inputs have often been used to allow evaluation of various summariser configurations, and can also provide end-users with the ability to control a summariser in an interactive manner.

3. DYNAMIC DOCUMENT SUMMARISATION

Once a range of summariser options can be specified interactively, users can iteratively manipulate inputs to tailor the resulting summary to their requirements. Such a system can be termed an interactive document summariser.

The TextSummary system [6] allows a user to provide parameters to the summarisation process, including summary length, specified as the number of sentences required in the summary. The IntelliScope Summariser [5] operates in a similar way, but allows users to specify a class of summary (such as 'outline' or 'executive'), and to control summary length and topic focus. HyperGen [17] produces interactive summaries, in the sense that non-extracted passages of text are accessible via hypertextual links. Link anchors contain labels to indicate the topic of the target text, and are interspersed between extracted sentences. WebSumm [12] extracts sentences from documents that have been returned by keyword queries, displaying them in a query result list. Sentences are chosen for extraction based on the user's query terms. The user can choose from a list of related terms to refocus the summaries, or expand the summaries using the current query terms.

A weakness of such systems is often the lack of responsiveness—users experience a delay between specifying options and presentation of the result. Consequently it is difficult for users to compare the effects of the changes that they make, and to rapidly investigate a range of settings. This problem is commonly experienced with query systems, and has resulted in the development of a new class of interface—dynamic query interfaces. These systems are characterised by immediate feedback to changes in query parameters, and the use of interface components such as slider bars to rapidly alter system settings. Research has shown this type is system to be supportive of users’ activities [7]. By applying these techniques to a document summariser one may support information providers in more efficiently producing semi-automatically generated summaries. End users will be able to more rapidly determine the utility of an on-line document by investigating a range of summary variants.

A number of tools exhibit some characteristics of such a *dynamic document summariser*. DataHammer [4] uses sentence extraction to summarise Web pages. As the user changes the value of a slider control, the summary immediately and progressively contracts or expands to correspond to the new compression level represented by the slider. The IntelliScope Summariser [5], Web Summariser [1] and Copernic Summarizer [3] provide similar facilities. To rapidly present an expanded or contracted summary the required processing of the document text must be minimised. Sentence extraction lends itself well to this situation because sentence scores are computed prior to presentation. When the required summary size changes, it is only necessary to compare sentence scores to the appropriate threshold value and display those sentences that are above the threshold. This computationally inexpensive approach enables immediate feedback.

A further characteristic of an interactive summariser is the ability to bias the summary towards a particular topic or set of topics.

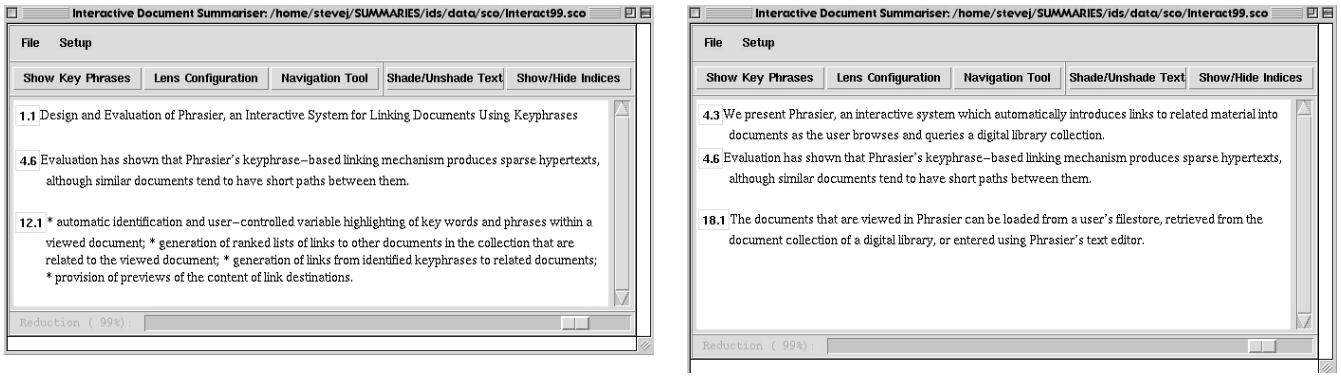

Figure 1: IDS document window (a) Standard summary at 99% compression (b) Summary at 99% compression with increased weighting of the topic 'digital library'.

This may be achieved in a number of ways. For example, the IntelliScope Summariser and the Web Summariser support summary focussing through specification of topic keywords.

A problem with creating a summary from extracted sentences is that at best, each sentence is lacking the context in which it originally occurred, and at worst a false context is implied to the reader. Natural Language Processing techniques may be used to ameliorate this problem (by resolving ambiguous references, for example), yet may conflict with the requirement for rapid feedback. DataHammer provides a basic indication of the context of extracted sentences via a visualisation tool. A graph shows each sentence in a document, and those that are extracted are highlighted. A reader can then tell, to some extent, whether summary sentences were contiguous in the source document, and where they appeared in the source.

4. IDS

The IDS system has been developed within the New Zealand Digital Library project [22], as part of our work on the development of user interfaces for effective access to on-line documents. IDS adopts the sentence extraction approach to the construction of summaries, and has two components. The first, written in Perl, contains tools to determine which sentences should be included in a summary. The second component, written in Tcl/Tk, is a user interface through which summaries can be tailored. First, sentence and paragraph boundaries are identified for each document, using a set of rules based on punctuation of the text. Next, keyphrases are automatically extracted from each document using the Kea keyphrasing system. Each sentence in each document is then awarded a score, using a heuristic based on the frequency of the document's keyphrases in the sentence. Sentences for each document can then be ranked in order of importance. The sentence and score data, and the keyphrase list are the inputs to the user interface.

4.1 Keyphrase Extraction

Keyphrases are extracted from a document using the Kea keyphrase extraction algorithm. Kea uses machine learning techniques to 'learn' what constitutes a good keyphrase. Kea has been described in detail elsewhere [9, 23] and we provide a summary here. There are two phases to Kea: learning a model of appropriate keyphrases, and use of the model to extract keyphrases from documents. To learn a model, Kea requires a set of training documents, for which there is a set of exemplar keyphrases (these might be provided by authors, or created by hand). A number of attributes of each acceptable keyphrase are determined in the context of all other candidate phrases in the training documents.

Once a model has been built it can be applied in the extraction stage, where new documents are processed. Candidate phrases from each document are tested against the model, and scored correspondingly. The higher the score, the more suitable the phrase is as a keyphrase of the document. The output for each document is a ranked list of keyphrases, their corresponding stems, and their scores. It is usual to apply some cut-off, either in the absolute number of phrases required, or the minimum acceptable score.

The model is a 'pluggable' component of the extraction process. Any Kea model can be applied to any document collection, although it clearly makes sense to apply a model derived from a related domain. Users will often bypass the model building stage and apply a pre-existing Kea model, such as *cstr*, a model derived from a set of computer science technical reports.

4.2 Sentence Scoring

First, sentence boundaries are identified within each document, using heuristics based on punctuation. Each sentence is allocated a unique identifier that denotes the paragraph in which it occurs, and its position within the paragraph. Each sentence from a given document is then scored in turn, based on the keyphrase stems that it contains. Stems are used so that strongly related phrases such as 'digital library', 'digital libraries', and 'digital librarian' are covered by a single stem 'dig libr'.

We have implemented a very simple scoring algorithm that arrives at a sentence score by summing the Kea scores for all stems that appear in the sentence. Sentences can then be ranked based on their score, with higher scoring sentences deemed to be the most suitable for inclusion in a summary of the document. This algorithm clearly favours longer sentences because they are more likely to contain more keyphrase stems. We believe that this is useful for a summary, where longer sentences tend to be more easily interpreted without surrounding context. Indeed, later in this paper, we show that this simple—and computationally inexpensive—algorithm produces better summaries than normative baseline approaches.

Once the scores have been calculated they are stored to disk and associated with the document. Sentence scoring is a one-off process—there is no need to repeat it when a document is accessed—IDS loads the sentence scoring information along with the document.

4.3 Controlling Summary Length

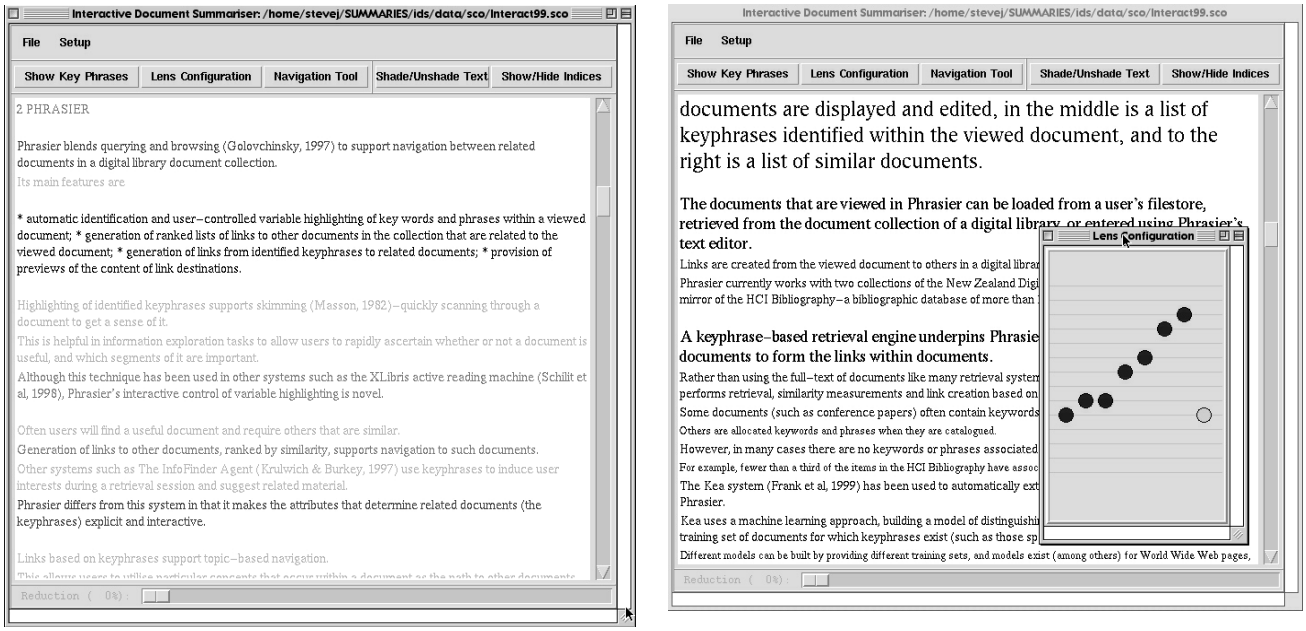

Figure 2: Summary in context views (a) sentence weights are mapped to a grayscale range for text shading (b) text is scaled based on sentence scores—the magnification lens can be interactively configured

The text of a document accessed through IDS is displayed in the document frame of the user interface. Figures 1 through 3 show a range of states of the document frame. This is currently a simple text viewer, but alternative presentations such as HTML are not difficult to provide. Paragraph boundaries from the original document are indicated by blank lines.

To the bottom of the document frame is a slider control that allows the user to dynamically alter the level of summarisation applied to the document. This level is initially zero, but as the user drags the slider to the right the summarisation level is increased, and the document text contracts dynamically. Figures 1(a) and 1(b) show summaries that are roughly 1% of the length of the original document. Here the slider is almost at the extreme right, and a compression level of 99% is shown. For the document in question, this produces a summary which is three sentences long.

The value (level of summarisation) of the slider control is mapped to the scores of sentences. As the value increases, the mapping determines the score threshold below which sentences should not be displayed. Any sentences with a score below this threshold are removed from the document viewer. The converse is true as the value decreases. The maximum value of the slider is 100, which results in the display of only the sentences with the highest score.

There are a number of possible alternative mappings between the slider and the score of sentences. For example, the slider value might indicate the percentage of sentences to be removed from the document to produce a summary. Thus a summarisation level of 33% would indicate that a third of the document has been removed and two thirds remain (the top 67% of the ranked list of sentences). Another option is to relate the slider value to the range of sentence scores for the document. In this case, a slider value of 50% might indicate that only sentences with scores above the median would be included.

Interactive control over the number of extracted sentences allows a user to move smoothly from summary to full text of the document. However, cues must be provided to enable a user to relate different summarisation levels.

4.4 Summary in Context

When a collection provider produces a summary, or a user reads a summary, it is possible that they will also access the full document. Consequently, the transition between abridged and full text should be supported, so that content and context of the summary is evident. A simple technique, shown in Figure 1, is to mark sentences in the summary with indices that describe their location in the full document. Each index shows the number of the paragraph containing a sentence, and the location of the sentence within the paragraph. Although this facility can quickly reveal which parts of the text have been removed, it does not reveal *what* has been removed.

Should a user consider a document to be of interest, having perused its summary, they may wish to view the entire document. The question arises as to how they might find the summary topics in the full text, or the surrounding context for the sentences that formed the summary. Location indices provide some indication as to where to look, but do not support a fluid transition from summary to document. IDS therefore provides *summary-in-context* views of the document.

One summary-in-context view uses text shading to reveal the summary sentences yet retain the context in which they occur in the full document. In Figure 2(a) the sentence scores are mapped to a grayscale range—the higher the score, the darker the shade. Important sentences are very prominent, and intermediate levels of importance are reflected. A similar approach is used in XLibris [19] and Phrasier [13] to highlight key words and phrases.

A further summary-in-context view uses text scaling, shown in Figure 2(b), to emphasise important sentences. Sentence scores are mapped to font size—the higher the score of a sentence, the larger the font used to display it. This is similar to the fish-eye text viewers described by Greenberg et al [11]. A user can set the differential between the text magnification levels to suit their preferences.

Each of the summary-in-context views—location indices, text shading and text scaling—can be combined, and applied in

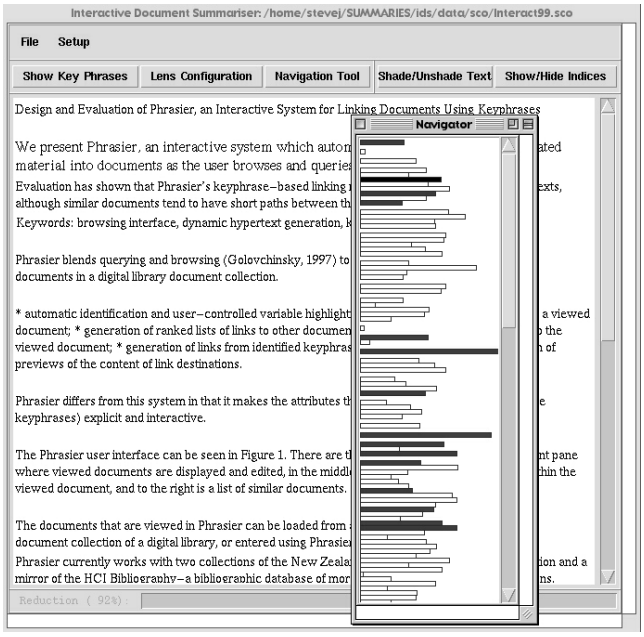

Figure 3: Summary context—the graphical overview reflects the locations in the source document from which summary sentences have been extracted.

conjunction with dynamic control over the summary length. Users are provided with a flexible range of presentations for the summary and its relationship with the full document.

IDS also provides a summary-in-context overview, in which a document map shows where extracted sentences occurred in the original document (shown in Figure 3). Sentences are represented by bars, with the first sentence at the top. The length of a bar represents sentence length, and each bar is shaded to reflect the importance of a sentence, using the same mapping as the text-shading view. Sentences that are displayed at the current summary level are highlighted. This view allows a user to quickly determine the extent to which each part of the source text is represented by summary sentences.

4.5 Focussing the Summary

When a collection provider or author produces a document summary, they may wish to focus it toward one or more aspects of the full document. IDS enables users to tailor not only the length of a summary, but also the topics of the document that are emphasised in the summary. This is achieved by amending the weights associated with the keyphrases that have been identified for the document. The keyphrases can be displayed in a ranked list, selected, and assigned new weights (expressed on a 0-100 scale). As soon as a new weight is confirmed, each sentence that contains the phrase in question is re-scored, and the display of the summary is updated accordingly. Such dynamic summary focussing can be rapidly achieved because we adopt the sentence extraction approach, and because the IDS sentence scoring algorithm is simple yet effective.

Figure 1 shows alternative three sentence summaries of the same document. To the left is the standard IDS summary. To the right is a summary produced by a simple topic weighting—the topic 'digital library' has been promoted, so that sentences that mention the topic are more likely to appear in the summary. Two of the sentences are now different. Note though that they also retain the theme of the original document, the Phrasier system.

Table 1: Number of sentences selected by subjects for each paper

Paper	Total Sentences	Mean Number Selected	SD	Mean selected (%)	SD (%)	% selected by at least 1 subject	% selected by at least 2 subjects
1	277	108	34	39	12	86	62
2	231	66	19	29	8	71	48
3	375	110	97	29	26	85	52
4	239	89	67	37	28	95	59
5	323	75	45	23	14	67	39
6	212	59	31	28	14	78	50

4.6 Anchoring Text Segments

Dynamic summary length and phrase weighting let users influence the final characteristics of a summary generated in a semi-automated manner. However, users may wish to manually specify that certain parts of the text are included in a summary. IDS therefore allows users to anchor sentences or paragraphs, so that they always appear in a summary, regardless of any other settings. The location indices shown in Figure 1 also act as anchor buttons. By clicking on a button, the user anchors the corresponding sentence, which is then displayed in a different colour from the rest of the text, to signify its anchored status. The user can anchor as many sentences and paragraphs as are required.

5. EVALUATION

As a first step in the evaluation of IDS we wished to investigate the quality of the summaries produced automatically by the system. There are a number of approaches to summary evaluation. One is to use corpora containing source documents and exemplar summaries, such as the TIPSTER materials used by Goldstein et al [10] and Lin [16]. Teufel and Moens [20] and Kupiec et al [15] used research papers with associated summaries provided by authors or professional summarisers. System performance can be measured by the similarity between the pre-existing and extracted summaries. Another approach, as used by Mitra et al [18], is to produce summaries for which human assessors then provide subjective judgements.

A problem in the evaluation of summaries produced by text extraction is that they are likely to be less readable than those produced by authors or professional abstractors. Consequently, negative subjective judgements might reflect summary characteristics other than the summariser's ability to extract the most appropriate sentences. A second problem is that existing summaries, such as abstracts, are unlikely to be formed simply from extracts of the source text. There may be no definitive measure of similarity between such abstracts and extracted summaries.

Therefore, we have designed an experiment in which human assessors carry out what is effectively sentence extraction from source texts. The performance of IDS has then been measured by its ability to identify and extract the sentences selected by people. Subjects were asked to consider a number of documents and to identify the sentences that conveyed the meaning of the document. The resulting sentences and inferred rankings were then compared to the output of IDS for the same documents. We wished to determine whether particular characteristics of the Kea keyphrases

Table 2: Distribution of selected sentences within each paper (values are rounded)

		% of selected sentences in segment s of each paper p						
		p1	p2	p3	p4	p5	p6	Mean
Segment	s1	7	10	11	11	8	10	10
	s2	10	13	9	9	9	8	10
	s3	15	10	8	10	9	4	9
	s4	13	11	11	9	13	7	11
	s5	12	12	12	10	9	6	10
	s6	9	8	13	10	9	11	10
	s7	8	9	11	8	17	9	10
	s8	7	13	10	11	5	11	10
	s9	11	8	9	12	7	16	11
	s10	10	7	7	10	14	17	11

could impact summarisation performance and so controlled them as independent variables of our evaluation.

5.1 Experiment Texts

A set of six English language papers from the Proceedings of ACM Conference on Human Factors 1997 (CHI 97; [2]) was used for the test documents. They provided a good fit with the background and experience of our subjects. Each paper was eight pages long.

5.2 Subjects

Subjects were recruited from a final year course on Human Computer Interaction being taken as part of an undergraduate degree programme in Computer Science. 18 subjects were recruited in all, of which 11 were male and seven female. All had completed at least three years of undergraduate education in computer science or a related discipline and were nearing completion of a fifteen week long course on human-computer interaction. The first language of each of the subjects was English. The youngest subject was 20, the oldest 49, and the mean age was 24

5.3 Method

The plain text of each paper was processed by Kea to extract keyphrases. A range of Kea parameters were controlled to produce a number of keyphrase lists for each document. Each of three keyphrase models was applied: *aliweb* is a model derived from a generic set of Web pages; *hcibib* is a model derived from bibliographic records (including abstracts) of Human Computer Interaction publications; and *cstr* is derived from a collection of Computer Science technical reports. Minimum and maximum keyphrase length was also controlled, producing list of phrases of 1-3 and 2-3 words in length. The number of keyphrases to be extracted from each document was also varied, with settings of five, 10 and 50 keyphrases. Therefore, given three models, two phrase lengths and three lists sizes, 18 phrase lists were produced for each document.

Each paper and phrase list was processed by IDS to extract and score sentences, giving a total of 18 alternative sentence rankings for each paper.

A baseline sentence ranking was required, against which to compare the performance of IDS. As with Lin [16] and Mitra et al [18] we considered the baseline sentence ranking to be the order in which they originally occurred in the source document. We believe this to be a stringent benchmark against which to measure IDS. Indeed, Mitra et al [18] go so far as to suggest that the quality of summaries formed by extracting lead segments from articles is good enough to bring into question the utility of text extraction techniques. Clearly, this may only hold for particular types of document. Novels, for example, are unlikely to be summarised successfully by selecting their initial sentences or paragraphs. However, newswire articles and documents such as the research papers that we have used already contain a great deal of leading summary information, such as title, abstract, introduction and so on.

Table 3: Agreement on selected sentences between subjects

		% of sentences selected by N subjects for paper p						
		p1	p2	p3	p4	p5	p6	Mean
N subjects	0	14	29	5	15	33	22	19
	1	24	23	36	33	28	27	29
	2	19	18	19	29	17	27	22
	3	19	15	18	13	15	14	16
	4	10	11	15	6	5	6	9
	5	6	3	7	5	2	2	4
	6	7	1	0	0	0	1	2

Two documents were randomly allocated to each subject, though the number of viewings and presentation order of each document were controlled. Each document was considered by six different subjects, and for three of the subjects it was the first document to be seen, and for the remaining three it was the second document to be seen. During the experimental session each subject carried out two tasks—the same task applied to two different documents. Subjects were allowed as much time as required for each task, and all completed both tasks within a single two hour period.

Each subject was given a copy of the paper in its original printed form, a pencil and eraser, and the following instructions

> Read the document through once
>
> Imagine that you are highlighting the document for revision purposes. You may wish to re-read the document at a later date, and to be able to find the important parts at that time. With this goal in mind, reread the document underlining whichever sentences you think are important.
>
> Underline all sentences which in your opinion
>
> - contain important points within the context of the document
> - are critical in conveying the message of the document
> - could not be removed from the document without detracting from its message
>
> You are free to underline as few or as many sentences as you wish. You can underline as many consecutive sentences as you wish.

After all subjects had completed the tasks, the selected sentences were identified and recorded for each paper, along with the frequency with which they were selected.

6. RESULTS

6.1 Number of Selected Sentences

The mean number of sentences that were selected by a subject varied across papers, both in terms of the absolute number and

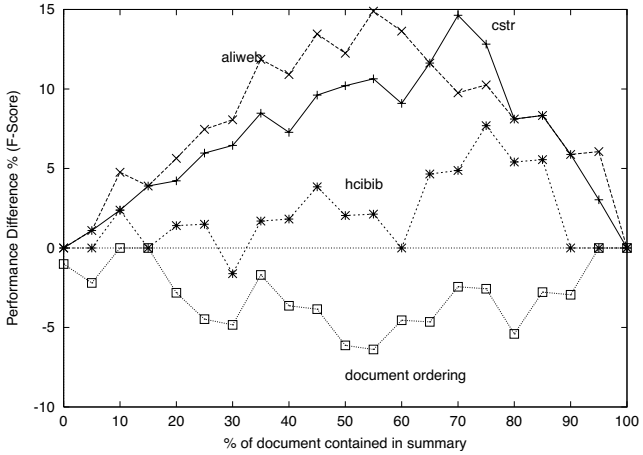

Figure 4: Comparative performance of the keyphrase extraction models

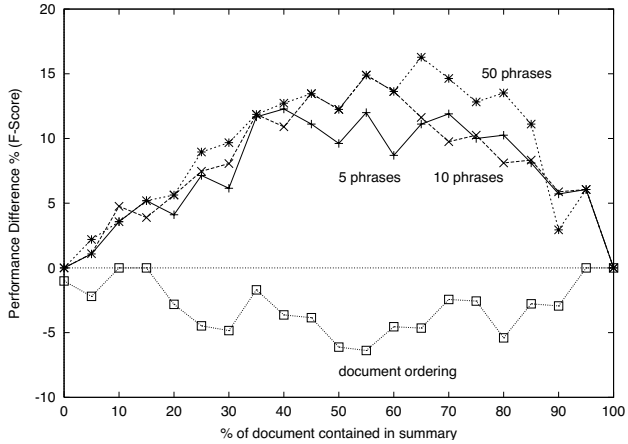

Figure 5: Comparative performance of sizes of keyphrase list extracted from source documents, using the aliweb model

proportion of sentences that were selected. From Table 1 we see that the smallest mean proportion selected was 23% for paper 5, and the largest 39% for paper 1. However there was substantial variation between subjects, as shown by the standard deviations.

In fact, one subject chose to select no sentences from either of the papers that s/he considered, although from observation it seemed that s/he gave the task appropriate and serious consideration. Another subject selected almost all of the sentences in the papers that s/he considered. Unfortunately we were unable to carry out follow up interviews with these subjects to determine their motivation for these extreme strategies. When we discard the data from these two subjects we find a rather more consistent picture, with a mean of roughly 30% of document sentences selected by a subject and a standard deviation of roughly 10% of document sentences.

When we consider selection across all subjects we see that the overwhelming majority of sentences in the documents were selected by at least one subject. When we look for agreement between at least two subjects on the selection of a sentence, we find that, on average, 50% of the sentences in the documents were selected.

6.2 Location of Selected Sentences

We also measured the distribution of selected sentences throughout each of the documents. Each document was divided into 10 segments of equal length and the number of sentences selected from within each segment was determined for all subjects. Table 2 shows selection patterns by paper and overall. The values show the proportion of selected sentences for each paper occurring within each segment. We observe that no particular document segment is favoured by the subjects. This is surprising, because a number of summarisation approaches consider location in the document to be an important attribute of a sentence [15, 16, 20]. Our observations indicate that this is not the case, and questions the utility of such an attribute. We might have expected the first segment of a document to be favoured because it contains both title and abstract, but it appears that subjects did not favour them over other parts of the text.

6.3 Length of Selected Sentences

We investigated whether subjects tended to choose longer sentences. Across each of the papers, the mean length of a sentence is 19.4 words (s.d.=11.9). This figure includes all document text, including section headings. The subjects tended to choose sentences that are slightly, but not significantly, longer than the mean. The mean length of a sentence selected by the subjects was 20.8 words (s.d.=11.6).

6.4 Subject Agreement

We next considered to what extent the sentence selection of the six subjects who read each paper overlapped. From Table 3, we see that on average, almost half of the sentences in a document were selected by zero or one subjects. Just over a fifth of the sentences were selected by two subjects. On average, only 15% of the sentences for a given paper were selected by a majority of the subjects.

When we consider individual papers, we see quite a variation from the mean scores. For example, 33% of the sentences in paper 5 were not selected, compared to only 5% of the sentences in paper 3. For paper 1, 7% of the sentences were selected by all six subjects, yet for each of the other papers zero (or almost zero) sentences were unanimously selected.

These observations are in keeping with previous findings that summaries created by humans, or human judgements of summaries tend to vary considerably. They serve to emphasise the need for a dynamic summariser which allows users to tailor a summary to their own requirements.

6.5 Performance of IDS

The central aspect of our evaluation focussed on how well the sentences selected by IDS matched those selected by human assessors. We have measured this in terms of precision and recall of 'relevant' sentences. The question arises as to how we might define a relevant sentence. Clearly, we can consider sentences that were selected by none of the subjects as 'irrelevant' or unsuitable for inclusion in a summary. We might then consider all sentences selected by at least one subject to be 'relevant'. However, this would not be a particularly stringent definition of relevance, and would not require agreement between subjects. Therefore, we have applied a definition of relevance that requires a sentence to have been selected by at least two subjects. We discuss below the implications of stricter definitions of relevance.

As our first performance measure we have computed F-scores, which reflect combined precision-recall performance at different summary sizes. The F-score is computed as (2 * precision * recall)/(precision + recall) and has been used by Lin [16] and Goldstein et al [10]. We computed F-scores for each set of

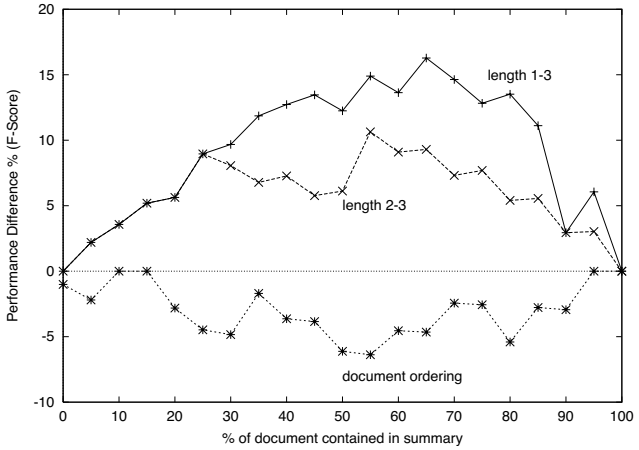

Figure 6: Comparative performance of different phrase lengths using the aliweb model

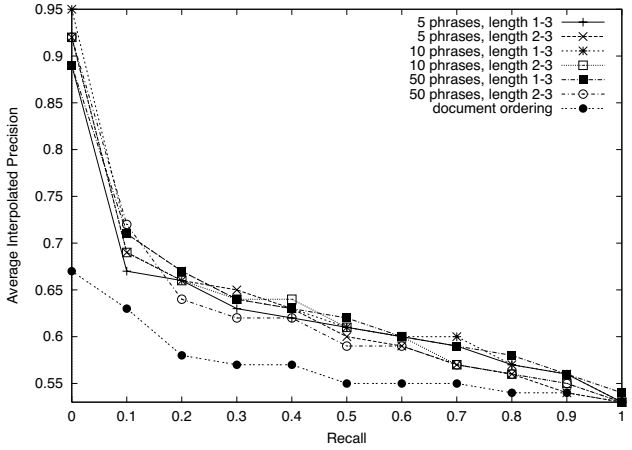

Figure 7: Precision-recall scores for alternative system settings using the aliweb model

Kea/IDS settings describe in Section 5.3, and for the baseline sentence ranking that matched the ordering of sentences in the source text. We also computed a baseline F-score as described by Goldstein et al [10]. This baseline measure reflects the degree to which the performance of a summariser can be accounted for by characteristics of the document itself and the chosen summary length. It accounts for the density of relevant sentences in a document (the higher the density, the higher the likelihood that a summariser will select relevant sentences), and the number of sentences selected.

For our second performance measure we have computed 11-point interpolated precision-recall curves. Here we take the list of extracted sentences ranked by IDS, and at each summary length determine the proportion of the included sentences that are relevant (precision) and the proportion of all relevant sentences that are included (recall). Precision is calculated at each of 11 values of recall.

6.5.1 Keyphrase Extraction Model

We first investigated whether performance differed for each of the Kea models used to extract keyphrases from the source documents. Figure 4 shows comparative performance values of IDS using each of the *cstr*, *hcibib* and *aliweb* models. Performance is based on F-scores and the performance curves are shown relative to the baseline achievable F-score. For example, for a summary that contains 25% of the sentences of the source text, *aliweb* is 7% better than the baseline, *cstr* 6% better, *hcibib* 1% better and the same ranking as the source text is 4% worse. This methodology follows that suggested by Goldstein et al [10].

The IDS summaries outperform baseline F-scores, and substantially outperform the simple ranking of sentences by the order that they appear in the text. Of the three Kea models used, *aliweb* produces the best summaries, followed by *cstr* and then by *hcibib*. Best performance is achieved when the summary consists of a substantial portion of the source text—around 55-60%.

6.5.2 Size of Document Keyphrase Set

Given the superior performance of the *aliweb* Kea model, we focussed on summaries produced using it to determine the extent to which the number of phrases extracted for a document affected performance. From Figure 5 we see that 50 phrases per document performed better than 10 phrases, which performed better than 5 phrases. However for summary lengths up to 50% of the original size of the source text, the performance improvement gained by using a larger number of phrases is minimal.

6.5.3 Phrase Length

Given that the *aliweb* model using 50 phrases produced the best performance improvements over the baseline F-score, we investigated the effect of phrase length under these conditions. From Figure 6 we see that phrases of between one and three words in length outperformed those consisting of only two or three words. However, for summary lengths up to 30% of the original source text, there is no difference in the gain in performance improvement.

6.5.4 Precision-Recall Curves

Figure 7 shows the 11-point interpolated precision-recall curves for summaries produced using the *aliweb* keyphrase model. The graph reflects the accuracy with which IDS includes relevant sentences, for proportions of the relevant sentences included in the summary. For example, when the summary contains 10% of the relevant sentences, roughly 70% of summary sentences are relevant, and 30% are not. As we normally observe with precision-recall curves, an increase in recall (more of the set of relevant sentences occur in the summary) results in a decrease in precision (the summary contains more non-relevant sentences).

The lower line on the graph reflects performance for a baseline summary formed by extracting sentences in the order in which they appear in the source text. The precision achieved by IDS is, on average, 16-19% better that that of the baseline summary.The curves for each of the phrase list size, and phrase length combinations are very similar. No combination clearly outperforms the others and the rates at which precision decreases are virtually identical.

Figure 8 plots the relationship between precision and summary length. The IDS data shown in the graph was generated using the *aliweb* model, producing 50 phrases of length 1-3 for each document. The baseline precision achieved by taking sentences in the order in which they occur in the document is also shown. Three definitions of a relevant sentence are shown—where a sentence was selected by one, two or three subjects. IDS produces higher precision than the baseline at all summary lengths and for each definition of relevance. Precision decreases slightly as summary length increases, and precision levels are substantially decreased when the documents are deemed to contain relatively few relevant sentences (more stringent definition of relevance).

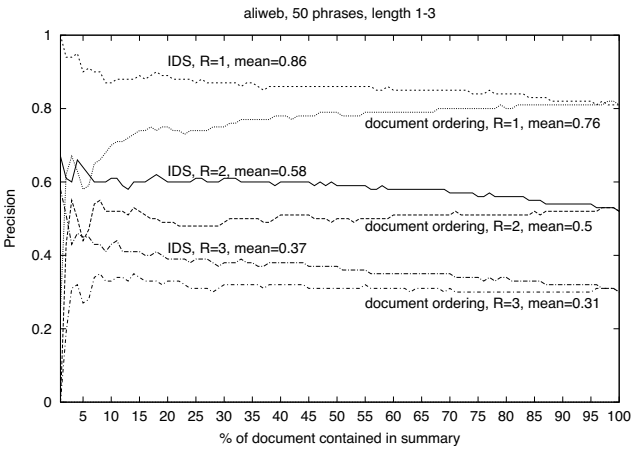

Figure 8: Precision at a range of summary lengths. R indicates the number of subjects selecting a sentence to deem it relevant.

6.5.5 Definition of Relevance

The F-scores and precision-recall measurements described above require a sentence to be selected by at least two subjects to be deemed relevant. Our notion of relevance can be made more stringent, to require selection by at least three subjects, by at least four subjects and so on. As we expect, with stronger notions of relevance the precision values and F-scores decrease.

However, we are interested in performance compared to the baseline sentence ranking, rather than absolute values. For stronger definitions of relevance we see the same relative performance as described above. The *aliweb* model provides best performance, then *cstr*, then *hcibib*. However, the disparity in performance between IDS summaries and those formed using document ordering of sentences decreases as fewer sentences in the documents are deemed to be relevant. When we define relevant sentences as those selected by at least five subjects we see summaries based on document ordering of sentences outperforming *hcibib* summaries. When we define relevant sentences as those selected by all six subjects to be relevant, document ordering outperforms all IDS summaries.

In the data that we report above, we find the greatest disparity between IDS summary performance and either the baseline F-score, or the document ordering of sentences at low compression levels of the source text, roughly 30-50%. However, as the relevance threshold is increased, the greatest disparity moves towards a compression level of 70%.

7. DISCUSSION

As with previous studies we observe substantial variation between the candidate extracts selected by our subjects. This comes as little surprise given the highly subjective nature of the task. It does, however, confirm our belief that summaries must be tailorable by both information providers and users (in a dynamic way) because there is no such thing as a single correct summary for a given document.

Sentences selected by subjects were distributed evenly throughout the source texts, which casts doubt upon the suitability of location as an indicator of the suitability of a sentence for extraction. For increasingly stringent definitions of relevance, where we require multiple subjects to select a sentence, we observed an increased tendency for relevant sentences to occur toward the start of the documents. However, the trend was not strong—perhaps due to the differences in structure and style between the papers.

On average, a single subject selected between roughly 20% and 40% of the sentences of a document, although there was substantial inter-subject variation. For the evaluation documents the standard summary (author's abstract) length is approximately 2% of the document text. This again implies that, for our test documents, a standard length summary will necessarily include only a small subset of all potentially appropriate sentences. Overall, subjects disagreed on which were the relevant sentences in a document, leading to high densities of selected sentences in each document. Once again this emphasises the need for flexibility in controlling summary content.

We thought that subjects may prefer longer sentences because of their ability to more fully present some idea. However, subjects tended to choose sentences that were little longer than the average length of sentences in the documents, and there was substantial variation in selected sentence length. We attribute this to the fact that selected sentences, even short ones, retained their context in the evaluation task, and so did not need to be 'self-contained'. The lack of observed preference for longer sentences calls into question extraction techniques which exploit sentence length as an attribute to direct extraction [15, 16].

Overall, the extracted IDS summaries gave substantially better performance than either simply taking sentences from the start of the document, or baseline F-scores which account for characteristics of the documents. It is encouraging that the IDS summaries perform well in comparison to baseline F-scores, suggesting that the use of automatically extracted keyphrases to identify summary sentences is effective.

Although Kea is a domain specific keyphrasing system we found, somewhat surprisingly, that the generic *aliweb* model extracted the keyphrases that led to the best summaries. On the surface, *hcibib* might be expected to have the best performance because of the match between the topic domain of the training documents and our evaluation texts. However, the style of documents is substantially different—the model was learned on bibliographic details not full research papers—and this appears to have been an important factor. The superior performance of *aliweb* is advantageous because it is generic, and can be widely applied to facilitate summarisation via keyphrases without the need to train further models.

We expected that by extracting larger numbers of keyphrases for the documents, we would improve the ability of IDS to select appropriate sentences. This was the case, to some extent, but only for longer summaries. It appears that for summaries up to 40% of the size of the source text, we need to extract only 5 Kea keyphrases from a document. This is beneficial because storage requirements are minimised, as are processing requirements—sentence scoring becomes faster as the number of document keyphrases decreases.

Our results show that the length of phrases extracted by Kea affects the quality of the extracted IDS summaries. Single-word keyphrases (such as 'library') tend to be generic when viewed in isolation, and may occur in multiple contexts. Multi-word keyphrases (such as 'digital library') can be far more descriptive. In fact, the quality of summaries up to 25% of the length of the source text was the same whether or not single word keyphrases were used. However, for longer summaries, single word keyphrases boost performance.

The precision values of Figure 8 compare favourably to those reported for other systems. For example, Teufel and Moens [20] also used, comparison of extracted sentences against those selected by a single assessor. They report a precision of 57.2% for their classification algorithm, using a combination of 5 sentence

characteristics, including occurrence of cue phrases. When we require at least two subjects to select a sentence for it to be relevant, IDS achieves mean precision of 58%. In fact, for very short summaries (2-5% of the document length) like those produced by Teufel and Moens, IDS achieves even better precision.

8. CONCLUSIONS AND FUTURE WORK

We have presented IDS, a document summarisation tool for maintainers and end users of digital libraries and similar systems. The IDS user interface provides novel, dynamic summarisation facilities that may help users to effectively judge the utility of on-line documents, and information providers to produce useful summaries in a semi-automatic way.

IDS uses a simple and efficient sentence scoring and extraction algorithm that supports the dynamic summary resizing and refocusing facilities provided through the user interface. Sentences are scored according to the occurence within them of keyphrases automatically extracted by the Kea system. Although simple, the algorithm produces summaries that are better than those produced by a common benchmark of selecting sentences from the start of a document, and performs as well as other sentence extraction systems.

Our evaluation methodology has directly addressed the measurement of the quality of sentences extracted for inclusion in summaries, by comparing them against those identified by human assessors. This approach minimises the limitations experienced by previous studies, such as readability and indirect mapping between extracted and 'gold-standard' summaries.

We have found that IDS produces its best summaries when using Kea keyphrases identified by a generic extraction model; that only a handful of keyphrases need be extracted for a document to be effectively summarised; and that single word keyphrases should be used.

Our initial results are encouraging, but we obviously wish to improve upon the current performance of IDS. One avenue for exploration is the sentence-scoring algorithm, which may be amended to use keyphrase rankings rather than scores. Another question is whether addition of author-keyphrases (where available) affects performance. These and other issues will be the focus of our future work.

9. References

[1] *Web Summariser*, Software Scientific Ltd, http://www.scientific.co.uk/software/htmlsum.htm

[2] *Proceedings of CHI'97: Human Factors in Computing Systems*, ACM Press, 1997.

[3] *Copernic Summarizer*, Copernic Technologies Inc., 2000, http://www.copernic.com/products/summarizer/

[4] *Data Hammer*, Glucose Development Corporation, 2000, http://www.glu.com/datahammer/

[5] *IntelliScope Document Summarizer*, Lernout & Hauspie, 2000, http://www.lhsl.com/tech/icm/retrieval/toolkit/ds.asp

[6] *The TextSummary Text Summariser*, 2000, http://www.textsummary.com

[7] Ahlberg, C., Williamson, C. and Shneiderman, B. Dynamic Queries for Information Exploration: an Implementation and Evaluation. In *Proceedings of CHI'92: Human Factors in Computing Systems*, (Moneterey, CA, 1992), ACM Press, 619-26.

[8] Delannoy, J.F., Barker, K., Copeck, T., Laplante, M., Matwin, S. and Szpakowicz, S. Flexible Summarization In *AAAI Spring Symposium Workshop on Intelligent Text Summarization*, (Stanford, CA, USA, 1998),,

[9] Frank, E., Paynter, G., Witten, I., Gutwin, C. and Nevill-Manning, C. Domain-specific Keyphrase Extraction. In *Proceedings of the Sixteenth International Joint Conference on Aritificial Intelligence*, 1999), Morgan-Kaufmann, 668-673.

[10] Goldstein, J., Kantrowitz, M., Mittal, V.O. and Carbonell, J.G. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In *Proceedings of SIGIR'99: the 22nd International Conference on Research and Development in Information Retrieval*, (Berkeley, CA, USA, 1999), ACM Press, 121-128.

[11] Greenberg, S., Gutwin, C. and Cockburn, A. Using Distortion-Oriented Displays to Support Workspace Awareness. In *People and Computers XI (Proceedings of HCI'96)*, 1996), Springer-Verlag, 299-314.

[12] House, D. *Interactive Text Summarization for Fast Answers*, The MITRE Advanced Technology Newsletter, 1997, http://www.mitre.org/pubs/edge/july_97/first.htm

[13] Jones, S. Design and Evaluation of Phrasier, an Interactive System for Linking Documents Using Keyphrases. In *Proceedings of Human-Computer Interaction: INTERACT'99*, (Edinburgh, UK, 1999), IOS Press, 483-490.

[14] Jones, S. and Paynter, G. Topic-based Browsing Within a Digital Library Using Keyphrases. In *Proceedings of Digital Libraries'99: The Fourth ACM Conference on Digital Libraries*, (Berkeley, CA, 1999), ACM Press, 114-121.

[15] Kupiec, J., Pedersen, J.O. and Chen, F. A Trainable Document Summarizer. In *Proceedings of SIGIR'95: the 18th International Conference on Research and Development in Information Retrieval*, (Seattle, Washington, USA, 1995), ACM Press, 68-73.

[16] Lin, C.-Y. Training a Selection Function for Extraction. In *Proceedings of the 8th International Conference on Knowledge Management*, (Kansas City, MO USA, 1999), ACM Press, 55-62.

[17] Mahesh, K. *HyperGen Summarization Tool*, 1997, clr.nmsu.edu/Research/Projects/minds/core_summarizer

[18] Mitra, M., Singhal, A. and Buckley, C. Automatic Text Summarization by Paragraph Extraction In *Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization*, (Madrid, Spain, 1997),,

[19] Schilit, B., Price, M. and Golovchinsky, G. Digital Library Information Appliances. In *Proceedings of Digital Libraries'98: The Third ACM Conference on Digital Libraries*, (Pittsburgh, PA, 1998), ACM Press, 217-226.

[20] Teufel, S. and Moens, M. Sentence Extraction as a Classification Task In *Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization*, (Madrid, Spain, 1997),,

[21] Turney, P.D. Learning Algorithms for Keyphrase Extraction. *Information Retrieval 2*, 4 (2000); 303-336.

[22] Witten, I.H., McNab, R.J., Jones, S., Apperley, M., Bainbridge, D. and Cunningham, S.J. Managing Complexity in a Distributed Digital Library. *IEEE Computer 32*, 2 (1999); 74-9.

[23] Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C. and Nevill-Manning, C.G. KEA: Practical Automatic Keyphrase Extraction. In *Proceedings of Digital Libraries '99: The Fourth ACM Conference on Digital Libraries*, (Berkeley, CA, 1999), ACM Press, 254-255.